

# Fitting a zeta distribution to a P survey: number of groups data

## Fitting a zeta distribution

Let us consider again the data from Roux et al. (2001). This data set is built into the package and can be accessed from the `Psurveys` object. That is, we can type:

```
> data("Psurveys")
> roux = Psurveys$roux
```

The package includes a special printing function that summarises the data for reading rather than displaying it in the way it is stored. R prints the values of objects, or variables, simply by typing their name. For example:

```
> roux
```

Number of Groups

n	rn
0	754
1	9
2	8
3	4
4	1

Roux C, Kirk R, Benson S, Van Haren T, Petterd C (2001).  
"Glass particles in footwear of members of the public in  
south-eastern Australia-a survey." *\_Forensic Science  
International\_*, \*116\*(2), 149-156.  
doi:10.1016/S0379-0738(00)00355-8  
<<https://doi.org/10.1016/S0379-0738%2800%2900355-8>>.

It is very simple to fit a zeta distribution to this data set. We do this using the `fitDist` function.

```
> fit = fitDist(roux)
```

We have assigned the result of the fitting to an arbitrarily chosen variable name, `fit`, chosen because it is easy to remember that it is a fitted object. The package includes specialised functions for both printing and plotting the fitted object. The `print` method displays an estimate of the shape parameter  $\alpha$  and the standard error of that estimate,  $\widehat{sd}(\hat{\alpha}) = se(\hat{\alpha})$ . The reported value is the same shape parameter used throughout `fitPS` and stored in the fitted object.

## Using the fitted distribution to estimate P terms

The `print` method displays the first 10 fitted probabilities from the model by default.

```
> fit
```

The estimated shape parameter is 4.9544  
The standard error of shape parameter is 0.2366  
The first 10 fitted values are:

P0	P1	P2	P3	P4
9.631547e-01	3.106447e-02	4.167082e-03	1.001917e-03	3.316637e-04

```

          P5          P6          P7          P8          P9
1.344002e-04 6.262053e-05 3.231467e-05 1.802885e-05 1.069709e-05

```

This information is probably sufficient for most casework. However, the package has a function, `probfun`, that returns a bespoke function that can calculate any probability term. This function is applied to a fitted object. For example:

```
> P = probfun(fit)
```

P is just a variable name and we could have used anything. We have chosen P because this probability function returns  $P$  terms. To use it, we only need to provide the value of  $k$ , and the function will return  $P_k$ . For example:

```
> P(5)
```

```

          P5
0.0001344002

```

## Fitting a zero-inflated zeta distribution

We can also fit a zero-inflated zeta model using the `fitZIDist` function. As before, we can choose a variable name to store the results in.

```
> fit.zi = fitZIDist(roux)
> fit.zi
```

The estimated mixing parameter, pi, is 0.8465

The estimated shape parameter is 2.8846

The first 10 fitted values are:

```

          P0          P1          P2          P3          P4
0.9716490911 0.0169404164 0.0052597614 0.0022938450 0.0012050764
          P5          P6          P7          P8          P9
0.0007122067 0.0004565511 0.0003106019 0.0002211302 0.0001631754

```

In the example above we fit a zero-inflated model to Roux et al.'s data and print the resulting fit. We get estimates of the parameters and a default set of fitted values. The output is interesting because the value of  $\hat{\pi}$  shows that the zero part of the zero-inflated model is picking up about 85% of the zeros. It is interesting to compare the estimates from the raw frequencies, the zeta model, and the ZIZ model. The estimates are shown in Table 1.

% latex table generated in R 4.5.3 by xtable 1.8-8 package % Wed Jun 10 17:21:41 2026

$k$	$P_k^{raw}$	$P_k^{zeta}$	$P_k^{ZIZ}$
0	0.9716	0.9632	0.9716
1	0.0116	0.0311	0.0169
2	0.0103	0.0042	0.0053
3	0.0052	0.0010	0.0023
4	0.0013	0.0003	0.0012
5	0.0000	0.0001	0.0007

Table 1: Estimated probability that  $k$  groups of glass would be found in shoes of a random member of the population based on the data of Roux et al. (2001), the raw frequencies, and those produced from the zeta and ZIZ models respectively.

We can see from Table 1 that we now have a non-zero estimate for  $P_5$ , but this comes at the cost of smaller probabilities for the preceding terms  $P_0$ – $P_4$ , which is not necessarily a negative. The survey data is dominated by zeros. However, we think it likely that the raw sample estimates for  $P_0$ – $P_4$  are overestimates. The model reduces the estimated value, which is in line with our thinking. Interestingly, the effect of including the

zero-inflation factor is to increase nearly all of the probabilities, with the exception of  $P_1$ . A natural question to ask is “Which model is correct?” The answer, unhelpfully, is “Neither”, because these are simply models. They can still help us without us having to believe that they are true.

## Confidence intervals for the parameter estimates

The `fitPS` package provides a `confint` method for the fitted value. The method returns both a Wald confidence interval and a profile likelihood interval. The two intervals are returned as elements of a `list` named `wald` and `prof`, respectively.

```
> ci = confint(fit)
> ci$wald
```

```
      2.5%      97.5%
4.490761 5.418099
```

```
> ci$prof
```

```
      2.5%      97.5%
4.520495 5.451277
```

## Bootstrapped and profile likelihood confidence regions for the zero-inflated zeta

The package includes the facility to compute both bootstrapped and profile likelihood confidence regions for the parameters of the zero-inflated zeta distribution. It also computes bootstrapped confidence intervals for the zeta distribution. The `confint` function returns a confidence region if the fitted object contains information from a zero-inflated zeta fit. As an example, we will first compute profile likelihood confidence regions for the Roux et al. (2001) data. To do this we use the fitted object we previously created, `fit.zi`, and, although not required, we supply a set of two levels so that we can compute both an 80% and a 95% confidence region. `confint` returns a list of confidence regions, one for each level, each of which is simply a set of  $x$  and  $y$  coordinates corresponding to the appropriate contour line.

```
> cr = confint(fit.zi, level = c(0.80, 0.95))
> plot(cr[["0.95"]], type = "l")
> polygon(cr[["0.8"]], border = "red")
> legend("topright", lty = 1, lwd = 2, col = c("red", "black"),
+        legend = c("80%", "95%"), bty = "n")
```

A bootstrapped confidence region can be computed using the `bootCI` function. The `bootCI` function includes the facility to plot the resulting confidence region or regions, and to hide or display the function’s progress. The latter is important because this procedure is numerically intensive and, even when utilising parallel processing, can be quite slow.

```
> bcr = bootCI(roux,
>             model = "ziz",
>             plot = TRUE,
>             silent = TRUE)
```

## Comparing two surveys

We can use the methodology that has been demonstrated so far to compare surveys. One reason for comparing surveys is to explore the hypothesis that there is no difference in the underlying true value of  $\alpha$ . If there is insufficient evidence to reject this hypothesis, then one may feel justified in combining data from two surveys. In the first instance we will take an ad hoc approach, and then treat this problem more formally. In our ad hoc approach we will compare confidence intervals for two surveys. If these confidence intervals overlap, then we might conclude that there is insufficient evidence in the data to suggest that the estimates of  $\alpha$  are different. We will illustrate this with the surveys conducted by Lau et al. (1997) and Jackson et

al. (2013). Lau et al. (1997) surveyed the clothing of 213 Canadian high school students and observed two sets of clothing with one fragment on each. Similarly, Jackson et al. (2013) surveyed 232 “randomly” selected members of the population of New South Wales in Australia. We place “randomly” in quotes because this was not a true random sample, but rather a convenience sample. That being said, it is unlikely that using a truly random mechanism would have significantly changed the results.

% latex table generated in R 4.5.3 by xtable 1.8-8 package % Wed Jun 10 17:21:41 2026

0	211	224
1	2	6

Table 2: Survey results from Lau et al. (1997) and Jackson et al. (2013).

Visual inspection of these surveys would suggest that they are fairly similar. We can fit a zeta distribution to each survey, and then compute a confidence interval for each survey. Again, these data sets are included in the `fitPS` package.

```
> lau = Psurveys$lau
> jackson = Psurveys$jackson
> fit.lau = fitDist(lau)
> fit.jackson = fitDist(jackson)
> confint(fit.lau)$wald
```

```
      2.5%      97.5%
4.948422 8.823417
```

```
> confint(fit.jackson)$wald
```

```
      2.5%      97.5%
4.443639 6.623473
```

We can see from the output that there is substantial overlap between these two Wald confidence intervals. The results using profile likelihood intervals lead to the same conclusion but are not shown. We can test this more formally. Specifically, we wish to test the null hypothesis that

$$H_0 : \alpha_1 = \alpha_2 \quad \text{or equivalently} \quad H_0 : \alpha_1 - \alpha_2 = 0,$$

where  $\alpha_1$  is the true value of  $\alpha$  for the Lau et al. data, and  $\alpha_2$  is the true value of  $\alpha$  for the Jackson et al. data. We choose a two-tailed alternative, meaning we are not concerned about the sign of any difference, but simply the magnitude of the difference. That is,

$$H_1 : \alpha_1 \neq \alpha_2 \quad \text{or equivalently} \quad H_1 : \alpha_1 - \alpha_2 \neq 0.$$

We test this hypothesis by constructing a test statistic and then computing a P-value under the assumption that the null hypothesis is true. We are interested in the difference between the two population values of  $\alpha$ . We estimate this by computing the difference in the sample estimates. That is, our estimate of  $\alpha_1 - \alpha_2$  is given by  $\hat{\alpha}_1 - \hat{\alpha}_2$ , where  $\hat{\alpha}_1$  and  $\hat{\alpha}_2$  are the maximum likelihood estimates based on the survey data. We scale this difference by the estimated standard deviation in the difference, that is, by the standard error of the difference,  $\text{se}(\hat{\alpha}_1 - \hat{\alpha}_2)$ . We estimate this—to keep the statistical theory to a minimum—as the square root of the sum of the two estimated variances, i.e.

$$\text{se}(\hat{\alpha}_1 - \hat{\alpha}_2) = \sqrt{\hat{V}(\hat{\alpha}_1) + \hat{V}(\hat{\alpha}_2)}.$$

Our test statistic is then

$$Z_0 = \frac{\hat{\alpha}_1 - \hat{\alpha}_2}{\text{se}(\hat{\alpha}_1 - \hat{\alpha}_2)}.$$

It can be shown that this test statistic follows an approximate normal distribution under the null hypothesis, which means our P-value can be computed by evaluating

$$P = \Pr(Z > |Z_0|) = 2(1 - \Pr(Z < |Z_0|)).$$

All this theory has been integrated into a function called `compareSurveys`.

```
> compareSurveys(lau, jackson)
```

Two-sided Wald test

```
data: lau and jackson
z = 1.1923, p-value = 0.2331
alternative hypothesis: true difference in shape parameters is not equal to 0
sample estimates:
  Shape of lau Shape of jackson
    6.885919      5.533556
```

The P-value is 0.23 (2 d.p.). This is significantly larger than either 0.05 or 0.01. Based on this, we would conclude that there is insufficient evidence to reject the null hypothesis of a common value of  $\alpha$ , and therefore it may be sensible to combine data from these two surveys. We could have also used the theory of likelihood ratio tests to test this hypothesis, but that is beyond the scope of this article. We note, however, that the `fitPS` package contains a function called `compareSurveysLRT` which can compare two or more surveys simultaneously using a likelihood ratio test.

## References

- Jackson, F., Maynard, P., Cavanagh-Steer, K., Dusting, T., and Roux, C. (2013). A survey of glass found on the headwear and head hair of a random population vs. people working with glass. *Forensic Science International*, 226(1), 125-131.
- Lau, L., Beveridge, A. D., Callowhill, B. C., Connors, N., Foster, K., Groves, R. J., Ohashi, K. N., Sumner, A. M., and Wong, H. (1997). The frequency of occurrence of paint and glass on the clothing of high school students. *Canadian Society of Forensic Science Journal*, 30(4), 233-240.
- Roux, C., Kirk, R., Benson, S., Van Haren, T., and Petterd, C. I. (2001). Glass particles in footwear of members of the public in south-eastern Australia: a survey. *Forensic Science International*, 116(2), 149-156.